DOCUMENT RESUME

ED 411 272 TM 027 244

AUTHOR Green, Donald Ross
TITLE Consequential Aspects of Achievement Tests: A Publisher's Point of View.
PUB DATE 1997-03-00
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Standardized Tests; State Programs; *Test Use; Test Validity; Testing Programs
IDENTIFIERS *Consequential Evaluation; *Test Publishers

ABSTRACT

It is argued that publishers of achievement tests, especially those who publish tests intended for use in many parts of the United States, are for the most part not in a position to obtain any decent evidence about the consequences of the uses that are made of their tests. What responsibilities and actions publishers can reasonably be expected to take, with respect to the consequences of test use, is explored. The uses of tests vary by teacher, school, district, state, and over time, especially the time between norming and test use, and no direct mechanism exists for obtaining evidence of the many consequences of test use. Publishers should undertake to study the matter for each of their tests to the extent possible and they should try to persuade academic researchers to study the matter objectively. Because there are so many tests and so many uses, it will take a large-scale cooperative effort to produce any generalizable evidence about the consequences of using nationally normed tests whatever their formats. This leadership cannot come from test publishers alone, but they should play a substantial role in the undertaking as they work with professional organizations to bring about many studies of test use. (Contains 12 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED411272

Consequential Aspects of Achievement Tests:

A Publisher's Point of View

By

Donald Ross Green

CTB/McGraw-Hill

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Donald Ross Green

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education

Chicago, IL, March 27, 1997

TM027244

BEST COPY AVAILABLE

The principal thesis of this paper is that the publishers of achievement tests, particularly the publishers of those tests intended for use in many districts across the nation, are for the most part not in a position to obtain any decent evidence about the consequences of the use that is made of their tests. After explaining the reasons why this is true, an attempt will be made to specify what responsibilities and actions publishers of such tests can reasonably be expected to take with respect to the many and varied consequences of use and what kinds of help must be obtained from the other parties in the enterprise.

The reasons why publishers of achievement tests are not typically able to investigate this aspect of validity are readily seen. The tests being discussed here are the familiar and widely used nationally normed and nationally marketed achievement batteries. Typically these tests are designed, developed and normed over a three or four year period and substantial use of them often is not made until some five or more years have passed from their conceptualization. The uses that are made of them are numerous and vary by teacher, by school, by district, by state, and over time.

Regardless of whether the statement of the construct being measured has been clearly stated by the publisher, each teacher, curriculum coordinator, test director, superintendent, school board member, state department official, state legislator, news reporter, and member of the various advisory and review committees has his or her own view of what constitutes reading, mathematics, science, social studies, and so forth. The responsibility for a clear and convincing description of the construct is obviously the responsibility of the publisher but any faith that such a statement has much to do with interpretations of the scores is ill founded. For example, some people seem to believe that all such tests are alike even though a conscientious review of these tests would almost surely disabuse the reviewer of that conviction.

No direct mechanism for obtaining really credible evidence of the many different consequences of the use of these achievement tests exists. Few if any schools or districts collect such evidence in a scientific manner. Furthermore, the typical school system uses a particular achievement test for about five years and then changes to a new test; consequently, by the time they may have accumulated evidence of the consequences of their use of the tests, they are no longer interested in that test.

Of course publishers do not operate completely blindly. They get customer complaints, they get customer questions – often about interpretations and uses. Some customers volunteer what they are doing with the tests and sometimes they make claims about good consequences; but what is really rare is solid scientific evidence. Publishers usually interview people and/or conduct focus groups of customers. However good these sources may be, the nature of these data on consequences for students is at best second or third hand, tends to be anecdotal, and is always hearsay.

Even the measurement community tends to rely on evidence of test consequences that is of this nature (e.g., Koretz et al 1994; Koretz et al 1996; Shepard, 1990). Look at the

claims that traditional NRTs narrow the curriculum and that therefore students learn less when they take such tests (Frederiksen, 1984, Madaus et al, 1992, Shepard, 1991). There is a visible likelihood that, in some instances at least, this is true based on the logic of the situation and reports of what teachers say they do (Madaus, 1992), but where is the experimental evidence? As far as I know, the attempt by Shepard et al (1996) is the first serious attempt to carry out such a study. Their findings do not show much support for the contention. While this result is probably for the reasons they offer, it remains undemonstrated.

Then there are the reports of the consequences of the "new sorts of assessments," namely performance assessments. For example, Kentucky has reported sharp gains on KIRIS and has suggested that this outcome arose because the testing program has led to better learning and instruction (Kentucky Department of Education, 1995). Perhaps the inference from the score gains is justified and I, for one, would certainly like to think so. But one cannot help noticing that a similar phenomenon, i.e., rising test scores, led to talk of "teaching to the test" and "the Lake Wobegon effect" just a few years ago when the tests in question were multiple-choice tests (Shepard, 1990, Phillips, 1990). It is also notable that CTBS scores, which had been rising when that test battery was the official evaluation of the Kentucky districts, stopped rising.

Now for some people, that merely indicates that what such multiple-choice achievement tests measure is irrelevant to "real learning." However, if that is the case, how does one explain that:

- as students go up the grades they score higher on such tests?
- generally acknowledged "good" students almost always score much higher on such tests than those not so acknowledged?
- teachers in the content area and grade rarely have difficulty with these multiple-choice tests?

One interpretation of these results is that the students in Kentucky, while maintaining their scores on CTBS, have not been able to generalize the greater knowledge and skill exhibited by the increase in scores on KIRIS. Obviously there are other possible interpretations. Since only some districts chose to give CTBS in those years, their uses varied and therefore both teacher concern and student motivation to perform varied.

A less striking, but possibly similar, result appears in Maryland, where the CTBS statewide scores clearly stopped rising when it was no longer the state test. The MSPAP scores went up the first year, but not the second, in reading, while in mathematics, somewhat lesser growth the first year was followed by a little further growth the second year. The variation from district to district in these patterns in both tests and their relation to each other is notable. Counter examples of almost any interpretation can be found in these data (Yen, 1996).

Thus, it appears that those who believe that performance assessment necessarily improves instruction have yet to make their case, and I believe that the data just cited opens to

question the assertions about the evils of multiple-choice tests.

Another common assertion is that multiple-choice tests encourage or even require the memorization of isolated facts and inhibit depth of conceptual learning and problem solving. It is hard to believe that many teachers allow this to happen, but perhaps it does in spite of the teachers. D'Ydewalle, Swerts and De Corte (1983) reported a study indicating that students told that they would be given an essay exam did better on a multiple-choice test than did those told that it was going to be multiple-choice partly because the latter group studied longer but apparently also because they studied differently. This finding lends support to the assertion. While Hakstian (1971) concluded that there was no such effect, Lundeberg and Fox (1991) point out that he did in fact report a significant difference favoring students with an essay set on multiple choice items measuring analysis. However Lundeberg and Fox conclude from their review of this matter that the data and research available are too thin to draw a conclusion.

The various studies and discussions of this issue suggest to me that students get impressions of what the tests are measuring mostly from teachers but also from each other. Neither of these categories of sources have ever nor will ever look at any publisher's statement of the construct being measured. In short this is an example how disconnected publishers are from the uses of their tests and why we cannot really respond well to the various public assertions about the consequences of uses of our tests.

Nevertheless publishers pay attention to these sorts of assertions about the consequences that flow from the use of the tests, especially when they are made by the widely quoted academic gurus who tend to say such things(the Bob Linns, the Pamela Mosses, and so forth). So this is a call for all those who believe they know or have better ideas about what tests and testing programs should be like to offer solid evidence about the consequences of the changes and improvements they are touting. For now, I submit that the "value implications" of these various score interpretations are inadequately evaluated (I am pretending that I understand Messick's facets (Messick, 1989)).

This is not to say that there is no reliable evidence about the consequences of using standardized achievement tests. Probably every publisher has some but I will limit myself to a few from CTB's experience. The "consequential aspects of validity" may be relatively new terminology but you will not be surprised to hear me say that concern about this matter is not new.

The very first thing I did when I came to CTB thirty years ago was to ask what uses were made of our tests, the California Achievement Tests in particular. The answer then was the same as it would be likely to be now from most people at CTB, to wit: the leading purpose of these achievement tests is "to help the teacher help the child." When asked how that was accomplished nobody seemed to really know. Therefore I went to a nearby school district and met individually with about ten elementary teachers in various grades and asked them what they did with the results. That particular investigation stopped there because none of them could name any concrete action they had taken from the data other

than using it to talk to parents (only a few of them did that.) I soon learned that many teachers did in fact use the results of the "Diagnosis of Learning Difficulties" based on a report showing right/wrong on items. Many of us were bothered by the unreliability of what were often single item scores; the ultimate upshot of all this was our move into criterion referenced tests which we began to publish in 1970.

Throughout much of the 1970s we conducted studies I dubbed "learner validation" studies most of which collected evidence about what happened when teachers used the results of these tests on an individual basis. The studies had many procedural and operational problems but generally they appeared to show that student achievement on tests specific to the objectives taught exhibited sharp gains in score. Did these programs help students? I truly believe they did but teachers found them complicated to implement and in the early '80s the interest in specific objectives began to fade. Consequently the necessary long-term follow up studies were never carried out and these sort of criterion-referenced tests disappeared much like their predecessors of the late 1920s.

What then can and should publishers do to meet their responsibilities? The options in order of increasing desirability and reasonableness are:

1. Ignore the issue and/or insist that it is entirely someone else's responsibility.
2. Undertake to seriously study the matter by themselves for each of their instruments.
3. Try to persuade academic researchers to study the matter objectively.
4. Try to work out some cooperative studies with individual customers.
5. Work through organizations such as NCME to get a series of systematic studies of the matter designed, financed and staffed involving many publishers, many school systems and many academics.

The merits and problems with most of these are numerous so only a few will be noted.

The first is clearly unacceptable to all of us or we would not be here even though it comes unfortunately close to representing the status quo.The second and third should perhaps be encouraged. A few serious studies might eventually appear in technical reports and a few others in journals some years later. However, the relevance of these reports to the testing programs then being set in place will almost always require generalizations well beyond the reported data and extrapolations to situations which differ in ways whose significance for the inferences is unknown, but the difficulties and limitations of one or a few isolated studies are legion. For example:

- Which users ? There are in the neighborhood of fifteen thousand school districts that use these tests.
- Which tests? The typical battery may have up to one hundred different tests; they cover anywhere from five to fifteen content areas in varying formats and differ substantially in content from grade to grade.
- Which uses? While there are probably not more ten or a dozen major uses of the scores, the variations in the way in which these are executed are quite large and no one knows which variation has which effect.

Given the number and range of such issues it seems self evident that only a large scale cooperative approach has any hope of shedding light on the general issue. Given that kind of cooperative effort, perhaps generalizable results might be possible for existing tests and some distinctions between the consequences of various kinds of achievement tests might be found.

The problems cited above for isolated studies are not necessarily solved by a cooperative effort and, of course, there are additional problems:

- Few school systems are likely to welcome reports of unanticipated negative consequences of their testing programs, so cooperation may be hard to obtain.
- Agreement among interested parties about the appropriate criterion measures of the consequences is likely to be contentious at best.
- Any cause-effect conclusions are likely to be disputed endlessly.
- If what has happened to date in the evaluation of performance assessment is any indication, much of the research undertaken is likely to be by those trying to prove that whatever exists is inferior to their new and better idea which of course will not be tested for many years.

I am sure that all of you can think of many more.

In short, given the circumstances that were described at the beginning of the paper it should be apparent that a huge set of studies would need to be done to yield substantial believable results applicable to more than a few of the uses made by a few of the customers of any one of the publishers of the nationally normed NRTs. I repeat that I believe it will take a large scale cooperative effort to produce any generalizable evidence about the consequences of the use of nationally normed tests whatever their formats. Because our business is highly competitive and extremely cost sensitive (dropping research studies is an easy way to cut costs), I do not believe that the leadership should come from the publishers alone but they should play a substantial role in the undertaking.

References

D'Ydewalle, G., Swerts, A., and De Corte, E. (1983). Study time and test performance as a function of test expectations. *Contemporary Educational Psychology, 8*(1), 55-67.

Frederiksen, N. (1984). The real test bias: Influences of testing and learning. *American Psychologist, 39*, 193-202.

Hakstian, A. R. (1971). The effects of type of examination anticipated on test preparation and performance. *Journal of Educational Research, 64*, 319-324.

Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final report: Perceived effects of the Maryland School Performance Assessment Program.* RAND.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*, 5-16.

Kentucky Department of Education. (1995, February). Celebrate the progress! 1992-94 Kentucky accountability results - summary news packet. Frankfort, KY: author.

Lundeberg, M.A., Fox, P.W. (1991) Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research* 61, 94-106.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York, NY: American Council on Education/Macmillan.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*, 15-26.

Shepard, L. A., Dougherty, K. C., (1991). Effects of high stakes testing on instruction. Paper presented at the annual meetings of the AERA and the NCME (Chicago, IL, April 3-7,1991).

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practice, 15*, 7-18.

Yen, W. M. (1996). *Linking Tests for AB 265.* Comparability Symposium. San Francisco.